

ITU-R

Radiocommunication Sector of ITU

Recommendation ITU-R BS.1116-3 (02/2015)

Methods for the subjective assessment of small impairments in audio systems

BS Series
Broadcasting service (sound)



Foreword

The role of the Radiocommunication Sector is to ensure the rational, equitable, efficient and economical use of the radio-frequency spectrum by all radiocommunication services, including satellite services, and carry out studies without limit of frequency range on the basis of which Recommendations are adopted.

The regulatory and policy functions of the Radiocommunication Sector are performed by World and Regional Radiocommunication Conferences and Radiocommunication Assemblies supported by Study Groups.

Policy on Intellectual Property Right (IPR)

ITU-R policy on IPR is described in the Common Patent Policy for ITU-T/ITU-R/ISO/IEC referenced in Annex 1 of Resolution ITU-R 1. Forms to be used for the submission of patent statements and licensing declarations by patent holders are available from <http://www.itu.int/ITU-R/go/patents/en> where the Guidelines for Implementation of the Common Patent Policy for ITU-T/ITU-R/ISO/IEC and the ITU-R patent information database can also be found.

Series of ITU-R Recommendations

(Also available online at <http://www.itu.int/publ/R-REC/en>)

Series	Title
BO	Satellite delivery
BR	Recording for production, archival and play-out; film for television
BS	Broadcasting service (sound)
BT	Broadcasting service (television)
F	Fixed service
M	Mobile, radiodetermination, amateur and related satellite services
P	Radiowave propagation
RA	Radio astronomy
RS	Remote sensing systems
S	Fixed-satellite service
SA	Space applications and meteorology
SF	Frequency sharing and coordination between fixed-satellite and fixed service systems
SM	Spectrum management
SNG	Satellite news gathering
TF	Time signals and frequency standards emissions
V	Vocabulary and related subjects

Note: This ITU-R Recommendation was approved in English under the procedure detailed in Resolution ITU-R 1.

Electronic Publication
Geneva, 2015

© ITU 2015

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without written permission of ITU.

RECOMMENDATION ITU-R BS.1116-3*

Methods for the subjective assessment of small impairments in audio systems

(Question ITU-R 62/6)

(1994-1997-2015)

Scope

This Recommendation is intended for use in the assessment of systems which introduce impairments so small as to be undetectable without rigorous control of the experimental conditions and appropriate statistical analysis. If used for systems that introduce relatively large and easily detectable impairments, it leads to excessive expenditure of time and effort and may also lead to less reliable results than a simpler test. This Recommendation forms the base reference for the other Recommendations, which may contain additional special conditions or relaxations of the requirements included in this Recommendation.

Keywords

Audio quality; small impairments; subjective assessment; listening test; audio coding; high-quality audio; listening room

The ITU Radiocommunication Assembly,

considering

- a) that Recommendations ITU-R BT.500, ITU-R BS.562, ITU-R BT.710 and ITU-R BT.811 have established some methods for assessing subjective quality of audio and video systems;
- b) that subjective listening tests permit assessment of the degree of annoyance caused to the listener by any impairment of the wanted signal during its transmission between the originating source and the listener;
- c) that classical objective methods may not be adequate in assessing advanced audio coding schemes and that perceptual objective assessment methods are being developed for testing the sound quality of sound systems;
- d) that the use of standardized methods is important for the exchange, compatibility and correct evaluation of the test data;
- e) that the introduction of new advanced digital audio systems exploiting psycho-acoustic properties, especially with small impairments requires advancements in subjective assessment methods;
- f) that the introduction of multichannel stereophonic sound systems up to 3/2 channels specified in Recommendation ITU-R BS.775 and the advanced sound system described in Recommendation ITU-R BS.2051, with or without accompanying picture requires new subjective assessment methods, including the experimental conditions,

recommends

- 1 that the testing, evaluation and reporting procedures given in Annex 1 be used for the subjective assessment of small impairments in audio systems including multichannel sound systems (with or without picture),

* This Recommendation should be brought to the attention of the International Organization for Standardization/Moving Picture Experts Group (ISO/MPEG) – Audio ad hoc Group.

further recommends

1 that further studies of the characteristics of listening rooms and reproduction devices for the advanced sound system are needed and this Recommendation should be updated when those studies are completed.

Annex 1

1 General

1.1 Contents

Annex 1 is divided into 11 sections, giving detailed requirements for various aspects of the tests:

1. General
2. Experimental design
3. Selection of listening panels
4. Test method
5. Attributes
6. Programme material
7. Reproduction devices
8. Listening conditions
9. Statistical analysis
10. Presentation of the results of the statistical analyses
11. Contents of test reports.

Also included are Attachments containing guidance on the selection of expert listeners and an example of the instructions given to the test subjects.

A number of common words are used with technical meanings. A Glossary of these is given in Attachment 4.

2 Experimental design

Many different kinds of research strategies are used in gathering reliable information in a domain of scientific interest. In subjective assessment of small impairments in audio systems, the most formal experimental methods shall be used. Subjective experiments are characterized firstly by actual control and manipulation of the experimental conditions, and secondly by quantitative data from human observers.

Careful experimental design and planning is needed to ensure that uncontrolled factors do not contaminate the listening test so that ambiguities are not caused. As an example, if the actual sequence of audio items is identical for all the subjects in a listening test, then one could not be sure whether the judgements made by the subjects were due to that sequence rather than to the different levels of impairments that were presented. Accordingly, the test conditions must be arranged in a way that reveals the effects of the independent factors, and only of these factors.

In situations where it can be expected that the potential impairments and other characteristics will be distributed homogeneously throughout the listening test, a true randomization can be applied to the presentation of the test conditions.

Where non-homogeneity is expected this must be taken into account in the presentation of the test conditions. For example, where material to be assessed varies in level of difficulty, the order of presentation of stimuli must be distributed randomly, both within and between sessions.

Similarly, listening tests need to be designed so that subjects are not overloaded to the point of lessened accuracy of judgement. Except in cases where the relationship between sound and vision is important, it is preferred that the assessment of audio systems is carried out without accompanying pictures.

A major consideration is the inclusion of appropriate control conditions. Typically, control conditions include the presentation of unimpaired audio materials, introduced in ways that are unpredictable to the subjects. It is the differences between judgement of these control stimuli and the potentially impaired ones that allows one to conclude that the grades are actual assessments of the impairments.

Some of these considerations will be discussed later in this document. It should be understood that the topics of experimental design, experimental execution, and statistical analysis are complex, and that only the most general guidelines can be given in a Recommendation such as this. It is recommended that professionals with expertise in experimental design and statistics should be consulted or brought in at the beginning of the planning for the listening test.

3 Selection of listening panels

3.1 Expert listeners

It is important that data from listening tests assessing small impairments in audio systems should come exclusively from subjects who have expertise in detecting these small impairments. The higher the quality reached by the systems to be tested, the more important it is to have expert listeners.

3.2 Criteria for selecting subjects

The outcome of subjective tests of sound systems with small impairments utilizing a selected group of listeners is not primarily intended for extrapolation to the general public. Normally the aim is to investigate whether a group of expert listeners, under certain conditions, are able to perceive relatively subtle degradations but also to produce a quantitative estimate of the introduced impairments. The demanding nature of the test procedure is intended to reveal those problems that may be revealed during the extensive period of exposure under different conditions which occur in real life once a system has been introduced to the consumer.

There is sometimes a reason for introducing a rejection technique either before (pre-screening) or after (post-screening) the real test. In some cases both types of rejection might be used. Here, elimination is referred to as a process where all judgements from a particular subject are omitted.

Any type of rejection technique which is not carefully analysed and applied may lead to a biased result. It is therefore extremely important that, whenever elimination of data has been made, the test report clearly describes the applied criterion so that the reader can make his own judgement.

3.2.1 Pre-screening of subjects

Pre-screening procedures, include methods such as audiometric tests, selection of subjects based on their previous experience and performance in previous tests and elimination of subjects based on a statistical analysis of pre-tests. The training procedure might be used as a tool for pre-screening.

The major argument for introducing a pre-screening technique is to increase the efficiency of the listening test. This must however be balanced against the risk of limiting the relevance of the result too much.

3.2.2 Post-screening of subjects

Post-screening methods can be roughly separated into at least two classes; one is based on inconsistencies compared with the mean result and another relies on the ability of the subject to make correct identifications. The first class is never justifiable. Whenever a subjective listening test is performed with the test method recommended here, the required information for the second class of post-screening is automatically available. A suggested statistical method for doing this is described in Attachment 1.

The methods are primarily used to eliminate subjects who cannot make the appropriate discriminations. The application of a post-screening method may clarify the tendencies in a test result. However, bearing in mind the variability of subjects' sensitivities to different artefacts, caution should be exercised.

3.3 Size of listening panel

The adequate size for a listening panel can be predicted if the variance can be estimated and the required resolution of the experiment is known.

Where the conditions of a listening test are tightly controlled on both the technical and behavioural side, experience has shown that data from 20 subjects is often sufficient for drawing appropriate conclusions from the test. If analysis of the data can be carried out as the test proceeds, then no further subjects need be processed when an adequate level of statistical significance for drawing appropriate conclusions from the test has been reached.

If some of the systems under test are expected to be nearly transparent, a larger number of subjects will be required to ensure that a sufficiently large number pass the post-screening test.

If, for any reason, tight experimental control cannot be achieved, then larger numbers of subjects might be needed to attain the required resolution.

The size of a listening panel is not solely a consideration of the desired resolution. The result from the type of experiment dealt with in this Recommendation is, in principle, only valid for precisely that group of expert listeners actually involved in the test. Thus, by increasing the size of the listening panel, the result can be claimed to hold for a more general group of expert listeners and may therefore sometimes be considered more convincing. The size of the listening panel may also need to be increased to allow for the probability that subjects vary in their sensitivity to different artefacts.

4 Test method

To conduct subjective assessments in the case of systems generating small impairments, it is necessary to select an appropriate method. The “double-blind triple-stimulus with hidden reference” method has been found to be especially sensitive, stable and to permit accurate detection of small impairments. Therefore, it should be used for this kind of test.

In the preferred and most sensitive form of this method, one subject at a time is involved and the selection of one of three stimuli (“A”, “B”, “C”) is at the discretion of this subject. The known reference is always available as stimulus “A”. The hidden reference and the object are simultaneously available but are “randomly” assigned to “B” and “C”, depending on the trial.

The subject is asked to assess the impairments on “B” compared to “A”, and “C” compared to “A”, according to the continuous five-grade impairment scale. One of the stimuli, “B” or “C”, should be indiscernible from stimulus “A”; the other one may reveal impairments. Any perceived differences between the reference and the other stimuli must be interpreted as an impairment.

As soon as the subject, in the preferred method, has completed the grading of a trial, it should be possible to proceed directly on to the next trial. The excerpt may be repeated until the subject has made an assessment. In this way the test procedure is self pacing.

The grading scale shall be treated as continuous with “anchors” derived from the ITU-R five-grade impairment scale given in Recommendation ITU-R BS.1284 and in Table 1.

TABLE 1

Impairment	Grade
Imperceptible	5.0
Perceptible, but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

NOTE 1 – It has been shown that the use of pre-defined intermediate anchor points may introduce bias [Poulton, 1992]. It is possible to use the number scales without descriptions of anchor points. In such cases, the intended orientation of the scales must be indicated. This may help to overcome translation problems in comparisons of tests carried out in different languages.

If intermediate anchor points are not used it is essential that the results for individual subjects are normalized with respect to mean and standard deviation. The following equation may be used to achieve such normalization whilst retaining the original scale.

$$Z_i = \frac{(x_i - x_{si})}{s_{si}} \cdot s_s + x_s$$

where:

- Z_i : normalized result
- x_i : score of subject i
- x_{si} : mean score for subject i in session s
- x_s : mean score of all subjects in session s
- s_s : standard deviation for all subjects in session s
- s_{si} : standard deviation for subject i in session s .

The use of scales without intermediate anchor points also precludes the interpretation of results in absolute terms.

It is recommended that the scale be used to a resolution of one decimal place.

The test method consists of two parts: a familiarization or training phase, and a grading phase.

4.1 Familiarization or training phase

Prior to formal grading, subjects must be allowed to become thoroughly familiar with the test facilities, the test environment, the grading process, the grading scales and the methods of their use. Subjects should also become thoroughly familiar with the artefacts under study. For the most sensitive tests they should be exposed to all the material they will be grading later in the formal grading sessions. During familiarization or training, subjects should be preferably together in groups (say, consisting of three subjects), so that they can interact freely and discuss the artefacts they detect with each other.

An example set of instructions is given in Attachment 3, “instructions to listeners”, as a model. Those instructions include a description of the “double-blind triple-stimulus with hidden reference” technique of stimulus presentation. Properly carried out, familiarization can transform some subjects with initially low ability into experts for the purposes of the test. By the end of the familiarization process, subjects should have arrived at a stable sense of the scale that will be used in the formal grading phase which will follow familiarization or training.

4.2 Grading phase

At the start of the first formal grading session of the day, an oral presentation of the test instructions should be made to each subject, preferably supplemented by written material. Several illustrative comparisons might be presented just before formal grading presentations are begun.

Since long- and medium-term aural memory is unreliable, the test procedure should rely exclusively on short-term memory. This is best done if a near-instantaneous switching (see Note 1) method is used in conjunction with a triple stimulus system as described in Attachment 3. Such switching demands close time alignment among the stimuli.

NOTE 1 – Exact instantaneous switching can produce artefacts if the waveforms of successive stimuli are not identical. For example, near-instantaneous switching with about 40 ms in total for fade-down/change-over/fade-up is preferred.

For the most critical assessments, one subject should be processed at a time. Only in this way can the subject exercise complete individual freedom to switch among the stimuli in the triple stimulus method. Such freedom is essential so that the subject can use his own discretion to fully explore the detailed comparisons among the stimuli of each trial.

Preferably, the subject should be able to switch between stimuli without visual guidance, so that, if the subject wishes, the eyes might remain closed for better concentration under conditions of minimal distraction. There should be no audible artefacts (e.g. “clicks”) of the switching system, since such artefacts can seriously interfere with the assessment process.

A grading session should not last for more than 20-30 min, although the self-paced character of trials advocated here will introduce uncontrolled variability among subjects. Experience suggests that no more than 10 to 15 trials per session should be scheduled to achieve the desired session length. Subject fatigue may become a major factor which would seriously interfere with the validity of judgements. To avoid this, rest periods equal to a duration no less than the session length should be scheduled between successive sessions for each subject.

5 Attributes

Listed below are attributes specific to monophonic, two-channel stereophonic and multichannel stereophonic (meaning up to 3/2 channels) and advanced sound system evaluations. It is preferred that the attribute “basic audio quality” is evaluated in each case. Experimenters may choose to define and evaluate other attributes.

The potential problem with having subjects try to assess more than one attribute on each trial is one of response burden. If subjects are overburdened or confused by trying to answer multiple questions about a given stimulus event, then this might produce unreliable gradings for all the questions.

5.1 Monophonic system

Basic audio quality

- This single, global attribute is used to judge any and all detected differences between the reference and the object.

5.2 Two-channel stereophonic system

Basic audio quality

- This single, global attribute is used to judge any and all detected differences between the reference and the object.

The following additional attribute may be of interest:

Stereophonic image quality

- This attribute is related to differences between the reference and the object in terms of sound image locations and sensations of depth and reality of the audio event.

Although some studies have shown that stereophonic image quality can be impaired, sufficient research has not yet been done to indicate whether a separate rating for stereophonic image quality as distinct from basic audio quality is warranted.

NOTE 1 – Up to 1993, most small impairment subjective evaluation studies of two-channel stereophonic systems have used the attribute basic audio quality exclusively. Thus the attribute stereophonic image quality was either implicitly or explicitly included within basic audio quality as a global attribute in those studies.

5.3 Multichannel stereophonic system

Basic audio quality

- This single, global attribute is used to judge any and all detected differences between the reference and the object.

The following additional attributes may be of interest:

Front image quality

- This attribute is related to the localization of the frontal sound sources. It includes stereophonic image quality and losses of definition.

Impression of surround quality

- This attribute is related to spatial impression, ambience, or special directional surround effects.

5.4 Advanced sound system

Basic audio quality

- This single, global attribute is used to judge any and all detected differences between the reference and the object. Consideration of attributes for advanced sound systems should be inclusive of attributes described for multichannel systems.

Additionally, the following attributes may be of interest:

Timbral quality – This attribute has been found to be of particular significance

- The attribute of timbral quality may be described by two sets of properties:
The first set of timbral properties is related to the *sound colour*, e.g. brightness, tone colour, coloration, clarity, hardness, equalization, or richness.
The second set of timbral properties is related to the *sound homogeneity*, e.g. stability, sharpness, realism, fidelity and dynamics. These properties may be descriptive of the timbre of the sound, but may also be descriptive of other characteristics of the sound.

Localization quality

- This attribute is related to the localization of all directional sound sources. It includes stereophonic image quality and losses of definition. This attribute can be separated into *horizontal localization quality*, *vertical localization quality* and *distant localization quality*. In case of the test with accompanying picture, these attributes can be also separated into *localization quality on the display* and *localization quality around the listener*.

Environment quality – This extends the attribute of surround quality

- This attribute is related to spatial impression, envelopment, ambience, diffusivity, or spatial directional surround effects. This attribute can be separated into *horizontal environment quality*, *vertical environment quality* and *distant environment quality*.

6 Programme material

Only critical material is to be used in order to reveal differences among systems under test. Critical material is that which stresses the systems under test. There is no universally “suitable” programme material that can be used to assess all systems under all conditions. Accordingly, critical programme material must be sought explicitly for each system to be tested in each experiment. The search for good material is usually time-consuming; however, unless truly critical material is found for each system, experiments will fail to reveal differences among systems and will be inconclusive.

It must be empirically and statistically shown that any failure to find differences among systems is not due to experimental insensitivity because of poor choices of audio material, or any other weak aspects of the experiment, before a “null” finding can be accepted as valid. In the extreme case where several or all systems are found to be fully transparent, then it may be necessary to program special trials with low or medium anchors for the explicit purpose of examining subject expertise (see Attachment 1).

These anchors must be known, (e.g. from previous research), to be detectable to expert listeners but not to inexperienced listeners. These anchors are introduced as test items to check not only for listener expertise but also for the sensitivity of all other aspects of the experimental situation.

If these anchors, either embedded unpredictably within the context of apparently transparent items or else in a separate test, are correctly identified by all listeners in a standard test method (see § 3 of this Annex) by applying the statistical considerations outlined in Attachment 1, this may be used as evidence that the listener’s expertise was acceptable and that there were no sensitivity problems in other aspects of the experimental situation. In this case, then, findings of apparent transparency by